SNS COLLEGE OF TECHNOLOGY

Coimbatore-37.

An Autonomous Institution

COURSE NAME : 19CSE301-INTRODUCTION TO DATA SCIENCE

III YEAR/ V SEMESTER

UNIT – V REPLICABILITY

Topic: Dimensionality Reduction

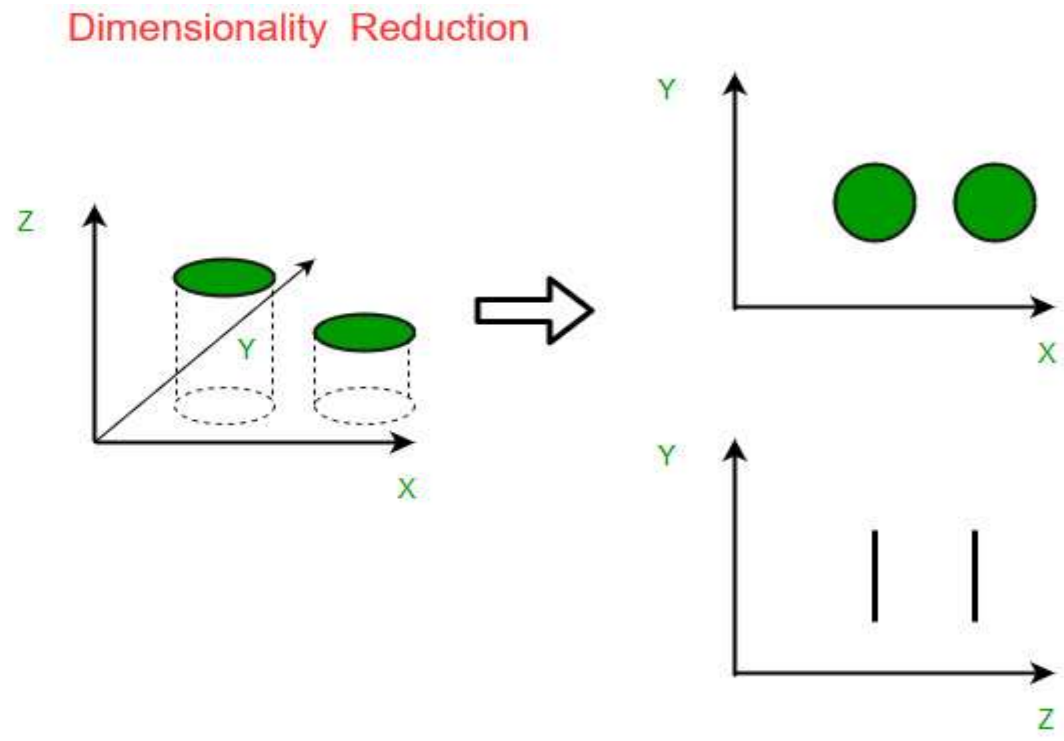
Mrs.G.Swathi
Assistant Professor
Department of Computer Science and Engineering

Introduction

- In machine learning classification problems, there are often too many factors on the basis of which the final classification is done.
- These factors are basically variables called features.
- The higher the number of features, the harder it gets to visualize the training set and then work on it
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

Introduction-Types

- It can be divided into feature selection and feature extraction.



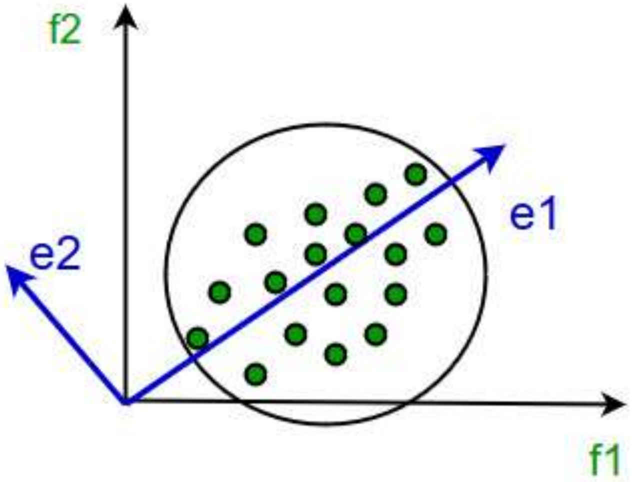
Components of Dimensionality Reduction

- **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  - Filter
  - Wrapper
  - Embedded
- **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Methods of Dimensionality Reduction

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

Dimensionality reduction may be both This method was introduced by Karl Pearson. It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum. linear or non-linear, depending upon the method used.



Steps & Advantage

- Construct the covariance matrix of the data.
- Compute the eigenvectors of this matrix.
- Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.

Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep- in practice, some thumb rules are applied

References

1 Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2013.

2 Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer; First Edition 2013.

3 P. Flach, ―Machine Learning: The art and science of algorithms that make sense of data, Cambridge University Press, 2012.

Thank you